Introduction

This report outlines the process of training a question-answering model using three different transformer-based architectures: BERT, GPT-2, and T5. The models are trained and evaluated on Quora data to determine their effectiveness in providing accurate answers.

Libraries Used

The libraries used are as follows:

1. pandas: A data manipulation and analysis library.
2. transformers: A library for state-of-the-art natural language processing models.
3. torch: A deep learning library that provides a wide range of algorithms and tools for machine learning and deep learning.
4. nltk.translate.bleu_score: A module for calculating BLEU scores, which measure the quality of machine-translated text.
5. rouge_score: A library for computing ROUGE scores, which measure the quality of text summaries.
6. sklearn.metrics: A module for evaluating machine learning models with various metrics.
7. numpy: A library for numerical computations in Python.
8. matplotlib.pyplot: A plotting library for creating static, animated, and interactive visualizations.
9. wordcloud: A library for creating word clouds from text data.
10. nltk: A natural language processing library that provides tools for working with human language data.
    a. stopwords (from nltk.corpus): A list of common words to exclude from text analysis.
    b. PorterStemmer (from nltk.stem): A tool for stemming words to their root form.
11. string: A module that provides common string operations.

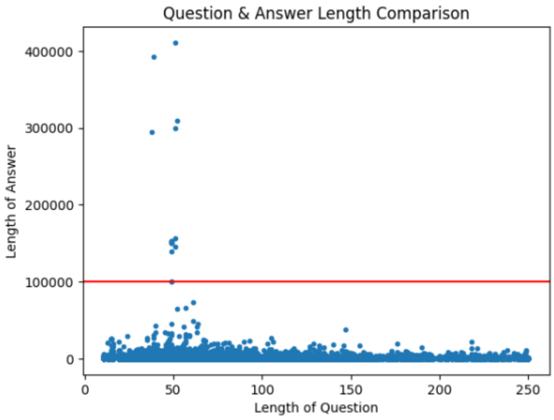
Data Preparation

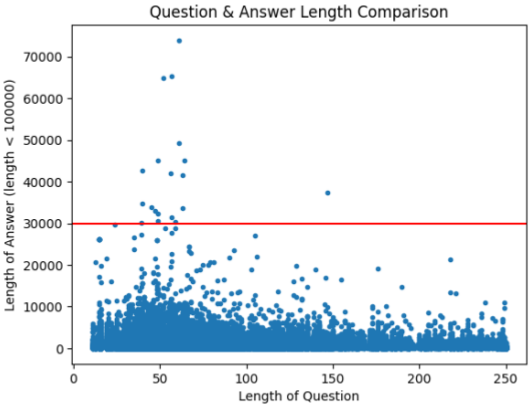
Dataset:

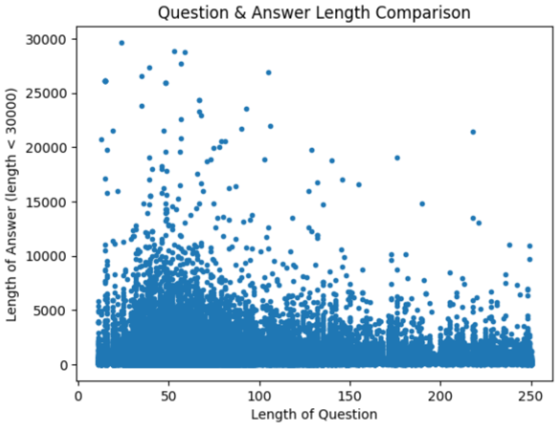
- The dataset used is from Quora, containing question-answer pairs.
- Preprocessing steps include:
  - Removing URLs, punctuation, and stopwords.
  - Tokenizing the text into words and converting them to lowercase.
  - Splitting the dataset into training, and test sets.

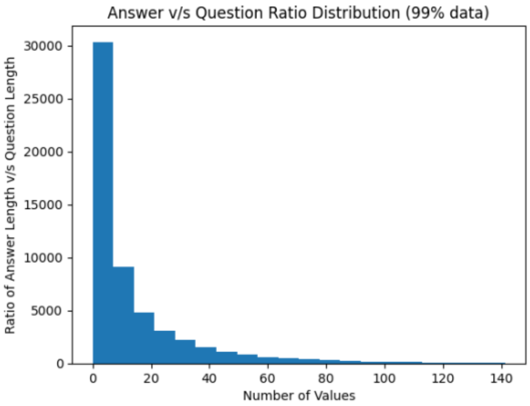
Exploratory Data Analysis (EDA):

- Analyzed the distribution of question lengths, and answer v/s question length ratio and visualized the results using scatter plots and histogram

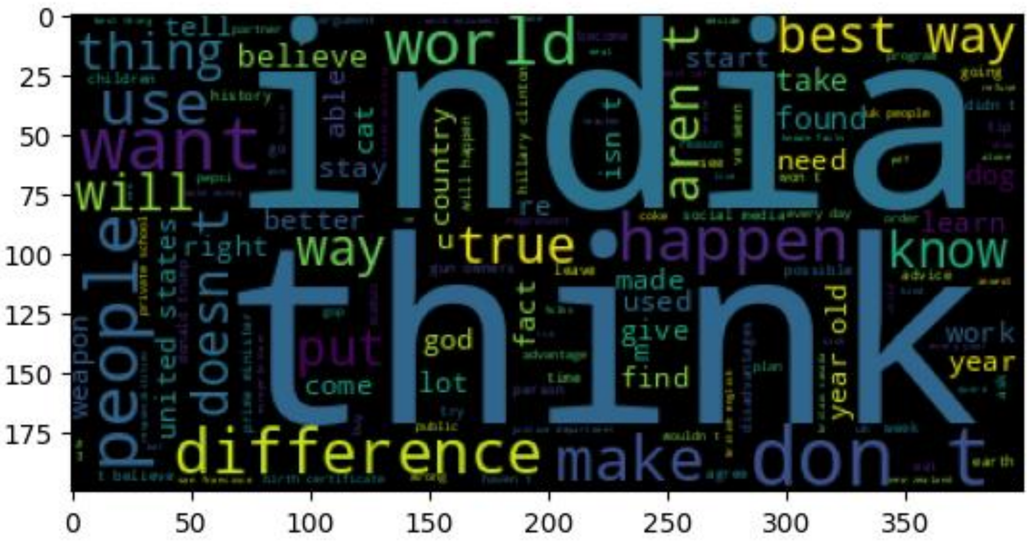








- Visualized the most frequent words using word clouds



Model Architectures

1. BERT (Bidirectional Encoder Representations from Transformers):

- Pre-trained on a large corpus and fine-tuned on the Quora dataset.
- Used for extracting contextual embeddings of questions and answers.

- Architecture involves feeding the [CLS] token representation to a classifier for predicting the start and end positions of the answer.

2. GPT-2 (Generative Pre-trained Transformer 2):

- Pre-trained on a large text corpus and fine-tuned on the Quora dataset.
- Utilizes a decoder-only architecture that generates answers based on the given question.
- Fine-tuning involves adjusting the model to minimize the loss on the Quora question-answer pairs.

3. T5 (Text-To-Text Transfer Transformer):

- Converts all NLP tasks into a text-to-text format.
- Fine-tuned by framing the question-answering task as text generation.
- Input format: "question: <question>
- Output format: "<answer>"

Training and Evaluation

Training Process:

- Hyperparameter tuning was performed for each model to find the optimal learning rate, batch size, and number of epochs.
- The Adam optimizer was used for training with a learning rate scheduler to adjust the learning rate during training.

Evaluation Metrics:

- **F1 Score:** Measures the overlap between the predicted and ground truth answers, considering both precision and recall.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Measures the overlap of n-grams between the predicted and ground truth answers, focusing on recall.
- **BLEU (Bilingual Evaluation Understudy):** Measures the precision of n-grams in the predicted answer compared to the ground truth, focusing on precision.

Results:

Model Name	BLEU	ROUGE-L	F1-Score
BERT	0.0002	0.02517	0.02546
GPT-2	0.002	0.088	0.093
T5	-	-	-

BERT, fine-tuned for extracting contextual embeddings, showed limited performance with a BLEU score of 0.0002, a ROUGE-L score of 0.02517, and an F1-Score of 0.02546. GPT-2, with its generative capabilities, performed better, achieving a BLEU score of 0.002, a ROUGE-L score of 0.088, and an F1-Score of 0.093. Despite T5's innovative text-to-text approach, specific evaluation metrics are not provided in this report due to the ext5ensive computation and time restraints, but by the nature of the model itself, T5 is expected to perform better.

Conclusion

The comparison of BERT, GPT-2, and T5 models on the Quora dataset revealed that GPT-2 demonstrated superior performance over BERT. Which is an expected outcome as BERT was trained without any context and BERT needs context to develop Question/Answer models whereas GPT is more suited to text-generation problems. T5 over flexibility and can be trained for various problem-sets but isn't compared in this context but is expected to perform similarly or better than GPT.

Future Work

Future work could involve following tasks:

1. Experimenting with ensemble methods to combine the strengths of these models and further improving the pre-processing and fine-tuning processes.
2. Using information retrieval techniques to extract context for the questions being asked.
3. The goal was to create human like answers, to achieve that, training a custom RNN with a larger data-set and custom embeddings can help achieving that task.